

A unified approach to Stein characterizations

Christophe Ley¹ and Yvik Swan²

*Department of Mathematics
Université Libre de Bruxelles
Campus Plaine – CP210
B 1050 Brussels, Belgium*

e-mail: chrisley@ulb.ac.be, yvswan@ulb.ac.be

Abstract: This article deals with Stein characterizations of probability distributions. We provide a general framework for interpreting these in terms of the parameters of the underlying distribution. In order to do so we introduce two concepts (a class of functions and an operator) which generalize those which were developed in the 70's by Charles Stein and Louis Chen for characterizing the Gaussian and the Poisson distributions. Our methodology (i) allows for writing many (if not all) known univariate Stein characterizations, (ii) permits to identify clearly minimal conditions under which these results hold and (iii) provides a straightforward tool for constructing new Stein characterizations. Our parametric interpretation of Stein characterizations also raises a number of questions which we outline at the end of the paper.

AMS 2000 subject classifications: Primary 20B10; secondary 62E17.

Keywords and phrases: Characterization theorems, Stein characterizations, location and scale parameters, parameter of interest, generalized (standardized) score function.

1. Introduction : background and motivation

Stein's method is a technique for obtaining bounds on a “distance” between an unknown probability distribution and a given *target distribution*. The method stems from two papers published in the 1970's by Charles Stein (concerning Gaussian approximation, see [33]) and Louis H. Chen (concerning Poisson approximation, see [9]). Since those days a substantial body of work has been devoted to extensions of the method, the literature on the subject now being vast and varied. We refer the reader to the monographs [4], [2], [3] or [11].

The gist of the method can be summarized as follows. Suppose that, for a given target distribution $g : \mathcal{X} \rightarrow \mathcal{X}$ dominated by a measure μ on some probability space $(\mathcal{X}, \mathcal{A}, \mu)$, there exists a class of functions $\mathcal{F}(g) \subset \mathcal{X}^* := \{\psi : \mathcal{X} \rightarrow \mathbb{R}, \mu - \text{measurable}\}$ and an operator $\mathcal{T}(\cdot, g) : \mathcal{X}^* \rightarrow \mathcal{X}^*$ such that

$$X \sim g(\cdot) \iff \mathbb{E}[\mathcal{T}(f, g)(X)] = 0 \text{ for all } f \in \mathcal{F}(g), \quad (1.1)$$

*This is an original survey paper

¹Supported by a Mandat d'Aspirant from the Fonds National de la Recherche Scientifique, Communauté française de Belgique. Christophe Ley is also member of E.C.A.R.E.S.

²Supported by a Mandat de Chargé de recherche from the Fonds National de la Recherche Scientifique, Communauté française de Belgique.

where by $X \sim g(\cdot)$ we mean $P(X \in A) = \int_A g(y) d\mu(y)$ for all measurable $A \subset \mathcal{X}$. Now let $Z \sim g(\cdot)$ and suppose that we are interested in studying a random object W whose distribution we do not know but which we believe to be approximately that of Z . After choosing a metric $d_{\mathcal{H}}(W, Z) := \sup_{h \in \mathcal{H}} |E[h(W)] - E[h(Z)]|$ for our approximation (where \mathcal{H} is also a certain class of functions), the first step in the Stein method consists in writing, for all $h \in \mathcal{H}$,

$$d_{\mathcal{H}}(W, Z) = \sup_{h \in \mathcal{H}} |E[\mathcal{T}(f_h, g)(W)]| \quad (1.2)$$

with f_h the solution of the so-called *Stein equation*

$$\mathcal{T}(f_h, g)(x) = h(x) - E[h(Z)]. \quad (1.3)$$

The intuitive reason for which (1.2) is an interesting quantity to study is the following: Z satisfies the rhs of (1.1), thus if the law of W is close to that of Z , then (1.1) should be nearly satisfied and the rhs of (1.2) should be close to 0 for all h such that $f_h \in \mathcal{H} \cap \mathcal{F}(g)$. Hence $|E[\mathcal{T}(f_h, g)(W)]|$ is an indicator of the \mathcal{H} -distance between W and Z .

The secret behind the method is that not only is the intuition outlined in the previous paragraph correct, but also, as it turns out, the rhs of (1.2) happens to be often “easier” to bound, making (1.2) a good starting point for a wide family of stochastic approximation problems. Determining equations of the form (1.1) for a given distribution g is the crucial starting of this method. For instance, Stein [33] showed that (1.1) holds for the Gaussian with $\mathcal{T}(f, g)(x) = f'(x) - xf(x)$ and $\mathcal{F}(g)$ the class of differentiable functions on \mathbb{R} ; Chen [9] showed a similar relationship for the rate- λ Poisson distribution, with $\mathcal{T}(f, g)(x) = f(x+1) - \lambda f(x)$ and $\mathcal{F}(g)$ the class of all bounded functions on \mathbb{Z} . After identifying a suitable characterization, the usual methodology relies on three steps, namely (i) solving (6.1) for all $h \in \mathcal{H}$, (ii) deriving bounds – the so-called *magic factors* – on the corresponding solutions, and (iii) applying the right tool (exchangeable pairs, zero- or size-biased distributions, truncation, etc.) in order to obtain explicit bounds on the rhs of (1.2) through the bounds obtained in step (ii). We refer the reader to the recent survey [31] for an overview.

This method has been applied in a wide number of problems. While the bulk of the (now vast) literature on this subject is devoted to Poisson and normal approximation problems, there have also been extensions towards other non standard densities, particularly so in recent years. Götze and Tikhomirov, for instance, use a characterization of the semi-circular law to obtain rates of convergence for spectra of random matrices with martingale structure (see [17]). Chatterjee, Fulman and Röllin use *two* different characterizations of the exponential distribution to obtain general results for convergence towards an exponential distribution (see [6]); they illustrate the applications of their methodology in a study of the spectrum of graphs with a non-normal limit. In [35], Stein, Diaconis, Holmes and Reinert obtain a characterization – by means of what is now called the *density approach* – of all regular distributions with a regular derivative (a function is regular if it is bounded and has at most countably many discontinuity points on its support, see also [34] and [12]); they use this in the analysis of

simulations. Eichelsbacher and Löwe [13] and Chatterjee and Shao [8] use [35]’s general characterizations to obtain general non-Gaussian approximation theorems, relevant for example in the field of statistical mechanics. Extensions to a multivariate setting are also available for the multivariate Gaussian law (see, for instance, [18], [7], or [30]). There exists a uniform treatment of the univariate discrete case, by means of Gibbs measures, which can be found in [14]. In [29] a characterization of the Kummer- U function is used to study degree asymptotics with rates for preferential attachment random graphs. Extensions of the method to continuous time processes are currently the object of active research (see [26], [28] or [27] and the references therein). This list is of course not exhaustive, and the method has also been used for binomial, negative binomial, multinomial, gamma, χ^2 and many other target distributions (see [11]).

In this paper we will not address Stein’s method *per se*, but rather concentrate on the characterizations (1.1) which are known, in the literature, as *Stein characterizations*. These have, so far, never been the subject of a specific treatment in the literature, and have always been introduced, through largely case-by-case arguments, as a means to an end rather than as an object of intrinsic interest. This is perhaps explained by the nature of the different target distributions (discrete, continuous, bounded or unbounded support, etc.) which make it complicated to try to unify all these characterizations under a single umbrella (in general different choices of target distributions will require imposing different combinations of restrictive assumptions). The purpose of this article is to exploit the similarities between *all* the characterizations discussed above in order to show how *all* these results can be seen as different instances of a unique phenomenon.

As it turns out, not only does our approach allow for (re-)obtaining the characterizations mentioned above (as well as many others) but it also simplifies the resulting proofs and allows to identify clearly the minimal conditions on the target densities under which such characterizations hold. More importantly it opens new lines of research, and builds hitherto unsuspected bridges between Stein’s method and information theory.

The outline of the paper is the following. In Section 2 we discuss different Stein characterizations and use this discussion to provide the heuristic behind our approach. This heuristic is formalized in Section 3 where we also prove our main result, Theorem 3.1, which provides a general and simple characterization theorem for a very broad family of (discrete and continuous, univariate and multivariate) distributions. In Section 4 we illustrate the consequences of Theorem 3.1 by providing characterizations for important classes of parametric distributions, namely the location and the scale families, as well as a characterization for discrete distributions. We apply our findings in a number of illustrative examples in Section 5, and uncover a couple of unpublished (to the best of our knowledge) characterizations. In Section 6, we discuss a couple potential applications of our results. Finally, Appendix A collects the more technical proofs.

2. Stein characterizations

In this section we provide the heuristic behind our approach. The arguments are constructive: starting from the Gaussian distribution we generalize so as to obtain the weakest possible assumptions for the most general result.

2.1. The density approach

Let ϕ stand for the standard Gaussian density. In a seminal paper [33], Charles Stein introduced the relationship

$$X \sim \mathcal{N}(0, 1) \iff \mathbb{E}[f'(X) - Xf(X)] = 0 \text{ for all } f \in \mathcal{F}(\phi), \quad (2.1)$$

with $\mathcal{F}(\phi)$ the collection of all differentiable real functions for which the expectation in (2.1) is defined. There exist many proofs of (2.1) (see, e.g., [19], [10] or [26]). We opt to present a different argument which enjoys the advantage of being transferable to virtually *any* continuous target density.

First remark that $-x = \phi'(x)/\phi(x)$ so that equation (2.1) can be equivalently rewritten

$$\begin{aligned} h(x) \propto \phi(x) &\iff \int_{\mathbb{R}} \left(f'(x) + \frac{\phi'(x)}{\phi(x)} f(x) \right) h(x) dx = 0 \text{ for all } f \in \mathcal{F}(\phi), \\ &\iff \int_{\mathbb{R}} \frac{(f\phi)'(x)}{\phi(x)} h(x) dx = 0 \text{ for all } f \in \mathcal{F}(\phi), \end{aligned} \quad (2.2)$$

where $h : \mathbb{R} \rightarrow \mathbb{R}^+$ is some density. The sufficient condition in (2.2) is immediate via integration by parts (the implicit boundary conditions on $f \in \mathcal{F}(\phi)$ ensuring that the constant term vanishes). To prove the necessity, choose for $A \subset \mathbb{R}$ a test function $f_A \in \mathcal{F}(\phi)$ that satisfies the differential equation

$$\frac{(f_A\phi)'(x)}{\phi(x)} = \mathbb{I}_A(x) - \int_A \phi(y) dy \quad (2.3)$$

for \mathbb{I}_A the indicator of A . If such a f_A exists and if it belongs to $\mathcal{F}(\phi)$, then (2.2) guarantees that

$$\int_A h(x) dx = \int_A \phi(x) dx \text{ for all } A \subset \mathbb{R}$$

and thus $h = \phi$. Of course (2.3) is easily solved explicitly, yielding the candidate solution

$$f_A(x) = \frac{1}{\phi(x)} \int_{-\infty}^x \phi(z) \left(\mathbb{I}_A(z) - \int_A \phi(y) dy \right) dz, \quad (2.4)$$

a function which, for all $A \subset \mathbb{R}$, is readily shown to satisfy all the requirements for belonging to $\mathcal{F}(\phi)$ (see, e.g., [10]). Hence the result holds. Now note that the above argument relies nowhere on specific properties of the Gaussian ϕ , but rather only on boundary and integrability conditions implicit in the definition of $\mathcal{F}(\phi)$ and in the solution (2.4). It therefore suffices to replace ϕ by some

generic density g in all the above arguments and to work out conditions on g so that everything runs smoothly in order to deduce, from (2.1), a general characterization theorem for continuous distributions.

To the best of our knowledge such a general characterization result was presented for the first time in [35] under a slightly different form; the only earlier similar attempt we have found is provided in [32] where a construction of Stein operators for Pearson and Ord families of distributions is provided. Stein's [35] result is now known in the literature as the *density approach* and allows for recovering the Gaussian characterization (2.1), the exponential characterization from [6] or the following two examples (which are also provided in [35]).

Example 2.1. Let $-\infty < a < b < \infty$. A random variable Z is $U[a, b]$ if and only if $E[f'(Z)] = f(b^-) - f(a^+)$ for all differentiable functions f .

Example 2.2. Let $\lambda > 0$. A random variable Z is $Exp(\lambda)$ if and only if $E[f'(Z) - \lambda f(Z)] = -\lambda f(0^+)$ for all differentiable functions f .

The denomination *density approach* is to be considered in analogy with the *generator approach* due to Barbour [1] and Götze [16].

2.2. Location-based and scale-based characterizations

There exist a number of outstanding characterizations which *cannot* be written in the form (2.2) such as e.g. those for discrete distributions or for the semi-circular distribution. For instance, in [6], a version of Stein's method for exponential approximation is developed, the arguments relying on two characterizations of the exponential distribution. The first, provided in Example 2.2, is an instance of the density approach. The second is given by

$$X \sim Exp(1) \iff E[Xf'(X) - (X-1)f(X)] = 0 \quad (2.5)$$

for all $f \in \mathcal{F}_2(Exp(1))$ a “sufficiently large class” of functions. This characterization is clearly not a consequence of the density approach. We nevertheless claim that (2.5) stems from the same origin as the characterization in Example 2.2; to see this it is necessary to re-interpret these results in terms of concepts inherited from a statistical point of view.

First recall how (2.2) was deduced by replacing the linear term x in (2.1) with the ratio $-g'/g$. This ratio is a familiar object in statistics: it is the *score function* $\varphi(x - \mu_0) = (\partial_\mu g(x - \mu)|_{\mu=\mu_0})/g(x - \mu_0)$, evaluated at $\mu_0 = 0$, associated with the *location* parameter μ of a location family $g(x - \mu)$ of distributions. Here ∂_μ stands for the derivative in the sense of distributions w.r.t. μ . With this parametric notation in hand, the characterization can be rewritten as

$$X \sim g(\cdot - \mu_0) \iff E \left[\frac{\partial_\mu (f(X - \mu)g(X - \mu))|_{\mu=\mu_0}}{g(X - \mu_0)} \right] = 0 \quad (2.6)$$

for all $f \in \mathcal{F}(g; \mu_0) (\supset \mathcal{F}(g))$ a sufficiently large class of functions depending on both g and μ_0 . This shows how the *density approach* can be seen as a spe-

cial instance (for $\mu_0 = 0$) of what we will henceforth call the *location-based characterization* (2.6).

Next reconsider equation (2.5). For a given $Exp(\sigma)$ distribution, the parameter σ is generally interpreted as a *scale* parameter. Writing out the argument of the expectation in the rhs of (2.6) in terms of a scale parameter of a scale family $\sigma g(\sigma x)$ of distributions leads to (∂_σ denotes the derivative in the sense of distributions w.r.t. σ)

$$\begin{aligned} \frac{\partial_\sigma(f(\sigma x)\sigma g(\sigma x))|_{\sigma=\sigma_0}}{\sigma_0 g(\sigma_0 x)} &= x f'(\sigma_0 x) + \frac{1}{\sigma_0} f(\sigma_0 x) + f(\sigma_0 x) x \frac{g'(\sigma_0 x)}{g(\sigma_0 x)} \\ &= x f'(\sigma_0 x) + f(\sigma_0 x) \left(\frac{1}{\sigma_0} + x \frac{g'(\sigma_0 x)}{g(\sigma_0 x)} \right). \end{aligned}$$

For g the density of an exponential distribution and $\sigma_0 = 1$, the latter equality corresponds to $x f'(x) + f(x)(1 - x)$, which is the argument of the expectation in (2.5) (note that for an exponential distribution, the support does not depend on the scale parameter, hence no indicator function needs to be differentiated). Thus, the second characterization of the $Exp(1)$ given in [6] can be viewed as a special instance (for $\sigma_0 = 1$) of what we will call a *scale-based characterization* which, in its most general form, reads

$$X \sim \sigma_0 g(\sigma_0 \cdot) \iff \mathbb{E} \left[\frac{\partial_\sigma(f(\sigma X)\sigma g(\sigma X))|_{\sigma=\sigma_0}}{\sigma_0 g(\sigma_0 X)} \right] = 0 \quad (2.7)$$

for all $f \in \mathcal{F}(g; \sigma_0)$ a sufficiently large class of functions depending on both g and σ_0 .

The location- and scale-based characterizations provided above do not, however, cover Chen's characterization of the Poisson distribution, to cite but this well-known example. Moreover, upon further thought, there is no intuitive justification which would explain why only location and scale parameters should play a special role; the tail parameter of a Student distribution or the upper and lower bounds of a uniform distribution over some interval $[a, b]$ should also be allowed to play a crucial role in such characterizations, as well as, e.g., the parameter λ of the Poisson distribution. As it turns out, there exists a much neater and efficient general framework in which both the above "general" results turn out to be straightforward particular cases.

2.3. A general characterization result

In this section we fix, for simplicity, $\mu_0 = 0$ and $\sigma_0 = 1$. Naïvely exploiting the similarities between (2.6) and (2.7) encourages us to propose the following general conjecture.

Conjecture 1. *Let $g(x; \theta)$ be a parametric family of densities with parameter θ . Suppose that $g(x; \theta)$ satisfies a number of regularity conditions. Fix a value θ_0*

of θ and denote by ∂_θ the derivative in the sense of distributions w.r.t. θ . Then

$$X \sim g(\cdot; \theta_0) \iff \mathbb{E} \left[\frac{\partial_\theta(f(X; \theta)g(X; \theta))|_{\theta=\theta_0}}{g(X; \theta_0)} \right] = 0 \quad (2.8)$$

for all $f \in \mathcal{F}(g; \theta_0)$ a sufficiently large class of functions depending on both g and θ_0 .

This Conjecture, if true, would enjoy several advantages: all kinds of parameters θ could appear, and no difference would be made between the continuous and the discrete case. While promising, the main drawback of (2.8) consists in the fact that the conditions on the target density g , as well as the structure of the family of test functions $\mathcal{F}(g; \theta_0)$ under which the Conjecture holds true, remain mysterious. In order to clarify this issue, one final argument needs to be invoked, the origin of which lies, once again, in a statistical approach to such identities.

Let $g(x; \theta)$ be as in the Conjecture above. A classical result in likelihood theory states that, under regularity conditions, the expectation of the score function $\partial_\theta g(x; \theta)/g(x; \theta)$ vanishes. The proof is very simple. Let $(\mathcal{X}, m_\mathcal{X})$ be a measure space (e.g., \mathbb{R} equipped with the Lebesgue measure or \mathbb{Z} equipped with the counting measure). Since $\int_\mathcal{X} g(x; \theta) dm_\mathcal{X}(x) = 1$, differentiating w.r.t. θ on both sides yields $\int_\mathcal{X} \partial_\theta g(x; \theta) dm_\mathcal{X}(x) = 0$, provided that the derivative and the integral are interchangeable. This immediately shows that the expectation under $g(x; \theta)$ of $\partial_\theta g(x; \theta)/g(x; \theta)$ equals zero. Now, under $g(\cdot; \theta_0)$, the rhs of (2.8) corresponds to $\int_\mathcal{X} \partial_\theta(f(x; \theta)g(x; \theta))|_{\theta=\theta_0} dm_\mathcal{X}(x) = 0$, which, under the condition of interchangeability of derivatives w.r.t. θ and integration w.r.t. x , can be rewritten as $\partial_\theta(\int_\mathcal{X} f(x; \theta)g(x; \theta) dm_\mathcal{X}(x))|_{\theta=\theta_0} = 0$. Thus, by analogy with the proof of the likelihood-based result, we see that, in order to belong to the class $\mathcal{F}(g; \theta_0)$, a test function f should satisfy the following natural three conditions in some neighborhood Θ_0 of θ_0 :

- (i) there exists a real constant c_f such that $\int_\mathcal{X} f(x; \theta)g(x; \theta) dm_\mathcal{X}(x) = c_f$ for all $\theta \in \Theta_0$;
- (ii) the mapping $\theta \mapsto \int_\mathcal{X} f(x; \theta)g(x; \theta) dm_\mathcal{X}(x)$ is differentiable over Θ_0 ;
- (iii) the differentiation w.r.t. θ and the integral sign are interchangeable for all $\theta \in \Theta_0$.

These conditions will be made more precise in Definition 3.1 of the next section. As we shall see, the first of these conditions yields the form of the candidate functions $f(x; \theta)$ (for instance $x \mapsto f(x - \theta)$ in the location case and $x \mapsto f(\theta x)$ in the scale case) and the second and third explain the sometimes complicated conditions imposed on the test functions in the relevant literature.

As a conclusion we stress an important fact: nowhere in the above argument do we rely on the target density to be continuous. As we will show in the following section, the heuristic outlined above holds irrespective of the nature of the target density, and (2.8) carries, as particular instances, the known characterizations for the Gaussian, the uniform, the exponential, the semi-circular, the Poisson and the geometric, to cite but these.

3. Characterizations in terms of a parameter of interest

In this section we present the main result of this paper, Theorem 3.1, which provides a unified framework for constructing Stein characterizations – by means of a characterizing class of test functions and a characterizing operator – for univariate, multivariate, discrete and continuous distributions. As announced in the previous section, we show that all these results allow for an interpretation in terms of a *parameter of interest* of the target distribution.

3.1. Notations and definitions

We first need to clearly identify the notations and vocabulary which will be used from now on. Throughout, we let $k, p \in \mathbb{N}_0$ and consider the two measure spaces $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, m_{\mathcal{X}})$ and $(\Theta, \mathcal{B}_{\Theta}, m_{\Theta})$, where \mathcal{X} is either \mathbb{R}^k or \mathbb{Z}^k , where Θ is a subset of \mathbb{R}^p whose interior is non-empty, where $m_{\mathcal{X}}$ is either the Lebesgue measure or the counting measure, depending on the nature of \mathcal{X} , where m_{Θ} is the Lebesgue measure, and where $\mathcal{B}_{\mathcal{X}}$ and \mathcal{B}_{Θ} are the corresponding σ -algebras. In this setup we disregard the case of discrete parameter spaces (as in, e.g., the discrete uniform); such distributions are shortly addressed in Remark 3.4 at the end of the current section.

Consider a couple (\mathcal{X}, Θ) equipped with the corresponding σ -algebras and measures. We say that the measurable function $g : \mathcal{X} \times \Theta \rightarrow \mathbb{R}^+$ forms a family of θ -parametric densities, denoted by $g(\cdot; \theta)$, if $\int_{\mathcal{X}} g(x; \theta) dm_{\mathcal{X}}(x) = 1$ for all $\theta \in \Theta$. In this case we call θ the parameter of interest for g . When $\mathcal{X} = \mathbb{R}^k$, corresponding to the absolutely continuous case, the mapping $x \mapsto g(x; \theta)$ is, for all $\theta \in \Theta$, a probability density function evaluated at the point $x \in \mathbb{R}^k$. When $\mathcal{X} = \mathbb{Z}^k$, corresponding to the discrete case, $g(x; \theta)$ is the probability mass associated with $x \in \mathbb{Z}^k$ and $g(\cdot; \theta)$ therefore maps \mathbb{Z}^k onto $[0, 1]$. This unified terminology will allow us to treat absolutely continuous and discrete distributions in one common framework. For the sake of simplicity, we rule out mixed distributions.

Example 3.1. θ -parametric densities are ubiquitous in probability and statistics. Taking $g : \mathbb{Z} \times \mathbb{R}_0^+ \rightarrow [0, 1] : (x, \lambda) \mapsto e^{-\lambda} \lambda^x / x! \mathbb{I}_{\mathbb{N}}(x)$, where $\mathbb{I}_A(\cdot)$ stands for the indicator function of some set $A \in \mathcal{B}_{\mathcal{X}}$, we see the density of a Poisson $\mathcal{P}(\lambda)$ distribution as a λ -parametric density. Taking $g : \mathbb{R} \times (\mathbb{R} \times \mathbb{R}_0^+) \rightarrow \mathbb{R}^+ : (x, (\mu, \sigma)') \mapsto (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/(2\sigma^2)}$, we see the density of a Gaussian $\mathcal{N}(\mu, \sigma)$ distribution as a (μ, σ) -parametric density. If, in the Gaussian case, the scale is known (and set to σ_0), one is then only interested in the location parameter μ . Taking $\tilde{g} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+ : (x, \mu) \mapsto \tilde{g}(x; \mu) = g(x; (\mu, \sigma_0)')$, we see the density of a Gaussian $\mathcal{N}(\mu, \sigma_0)$ distribution as a μ -parametric density. Likewise, one can see the density of a uniform $U[a, b]$ distribution as an (a, b) -parametric density, an a -parametric density or a b -parametric density. In general, there are infinitely many ways to write the density of any given probability distribution as a θ -parametric density for any given θ . See for instance, on this issue, the discussion on the so-called natural parameters of the exponential family in [24].

Fix a couple (\mathcal{X}, Θ) as above, endowed with their respective σ -algebras and measures. Throughout this paper, the densities we shall work with all belong to the class $\mathcal{G} := \mathcal{G}(\mathcal{X}, \Theta)$ of θ -parametric densities for which the mapping $\theta \mapsto g(\cdot; \theta)$ is differentiable in the sense of distributions. Such distributions may have a bounded support possibly depending on the parameter θ ; we will denote this support by $S_\theta := S_\theta(g)$, be it dependent on θ or not.

With this in hand, we are ready to define the two fundamental concepts of this paper. These (a class of functions and an operator) mirror notions already present in the literature on Stein characterizations.

Definition 3.1. *Let θ_0 be an interior point of Θ and let $g \in \mathcal{G}$. We define the class $\mathcal{F}(g; \theta_0)$ as the collection of test functions $f : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ such that the following three conditions are satisfied in some neighborhood $\Theta_0 \subset \Theta$ of θ_0 .*

Condition (i) : there exists $c_f \in \mathbb{R}$ such that $\int_{\mathcal{X}} f(x; \theta) g(x; \theta) dm_{\mathcal{X}}(x) = c_f$ for all $\theta \in \Theta_0$.

Condition (ii) : the mapping $\theta \mapsto f(\cdot; \theta) g(\cdot; \theta)$ is differentiable in the sense of distributions over Θ_0 .

Condition (iii) : there exist p $m_{\mathcal{X}}$ -integrable functions $h_i : \mathcal{X} \rightarrow \mathbb{R}^+$, $i = 1, \dots, p$, such that $|\partial_{\theta_i}(f(x; \theta) g(x; \theta))| \leq h_i(x)$ over \mathcal{X} for all $i = 1, \dots, p$ and for all $\theta \in \Theta_0$.

The three conditions in Definition 3.1 are to be compared to the three conditions discussed at the end of Section 2.3.

Definition 3.2. *Let θ_0 be an interior point of Θ . Also let g and $\mathcal{F}(g; \theta_0)$ be as above. We define the Stein operator $\mathcal{T}_{\theta_0} := \mathcal{T}_{\theta_0}(\cdot, g) : \mathcal{F}(g; \theta_0) \rightarrow \mathcal{X}^*$ as*

$$\mathcal{T}_{\theta_0}(f, g)(x) = \frac{\nabla_{\theta}(f(x; \theta) g(x; \theta))|_{\theta=\theta_0}}{g(x; \theta_0)}. \quad (3.1)$$

The operator defined by (3.1), inspired by the rhs of (2.8), requires some comments. If the support of $g(\cdot; \theta)$ is \mathcal{X} itself, then the operator is obviously well-defined everywhere. If, on the contrary, the density $g(\cdot; \theta)$ has support $S_\theta \subset \mathcal{X}$, then there is an ambiguity which we need to avoid. To this end we adopt the convention that, whenever an expression involves the division by an indicator function \mathbb{I}_A for some $A \in \mathcal{B}_{\mathcal{X}}$, we are, in fact, multiplying the expression by the said indicator function. With this convention, writing out the operator in full (whenever the gradient $\nabla_{\theta}(f(x; \theta))|_{\theta=\theta_0}$ is well-defined on \mathcal{X}) reads

$$\mathcal{T}_{\theta_0}(f, g)(x) = \left(\nabla_{\theta}(f(x; \theta))|_{\theta=\theta_0} + f(x; \theta_0) \frac{\nabla_{\theta}(g(x; \theta))|_{\theta=\theta_0}}{g(x; \theta_0)} \right) \mathbb{I}_{S_{\theta_0}}(x).$$

Our convention not only guarantees that the Stein operator is well-defined but also that, for any test function f , the support of $\mathcal{T}_{\theta_0}(f, g)(x)$ is included in S_{θ_0} . This convention was implicit throughout the discussion in the heuristic section. As already mentioned there, the usage of derivatives in the sense of distributions of g with respect to θ implies also taking derivatives of indicator functions whenever S_θ depends on θ .

Example 3.2. (i) Let $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$ and $g(x; \mu) = (2\pi)^{-1/2} e^{-(x-\mu)^2/2}$, the density of a univariate normal $\mathcal{N}(\mu, 1)$ distribution. Clearly, g belongs to \mathcal{G} for all $\mu \in \mathbb{R}$ and its support $S_\mu = \mathbb{R}$ is independent of μ . Fix $\mu_0 = 0$ and consider functions of the form $f : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} : (x, \mu) \mapsto f(x; \mu) := f_0(x - \mu)$, where $f_0 : \mathbb{R} \rightarrow \mathbb{R}$ is chosen such that $f \in \mathcal{F}(g; 0)$. Restricting the operator \mathcal{T}_0 to the collection of f 's of this form, it becomes

$$\mathcal{T}_0(f, g)(x) = -f'_0(x) + x f_0(x).$$

(ii) Let $\mathcal{X} = \mathbb{Z}$, $\Theta = \mathbb{R}_0^+$ and $g(x; \lambda) = e^{-\lambda} \lambda^x / x! \mathbb{I}_{\mathbb{N}}(x)$, the density of a Poisson $\mathcal{P}(\lambda)$ distribution. Clearly, g belongs to \mathcal{G} for all $\lambda \in \mathbb{R}_0^+$ and its support $S_\lambda = \mathbb{N}$ is independent of λ . Fix $\lambda = \lambda_0$ and consider functions of the form $f : \mathbb{Z} \times \mathbb{R}_0^+ \rightarrow \mathbb{R} : (x, \lambda) \mapsto f(x; \lambda) := e^\lambda [\lambda f_0(x+1)/(x+1) - f_0(x)]$, where $f_0 : \mathbb{Z} \rightarrow \mathbb{R}$ is chosen such that $f \in \mathcal{F}(g; \lambda_0)$. Restricting the operator \mathcal{T}_{λ_0} to the collection of f 's of this form, it becomes

$$\mathcal{T}_{\lambda_0}(f, g)(x) = e^{\lambda_0} \left(f_0(x+1) - \frac{x}{\lambda_0} f_0(x) \right) \mathbb{I}_{\mathbb{N}}(x).$$

Among densities $g \in \mathcal{G}$, those which satisfy the following (local) regularity assumption at a given interior point $\theta_0 \in \Theta$ will play a particular role.

Assumption A : there exists a rectangular bounded neighborhood $\Theta_0 \subset \Theta$ of θ_0 and a $m_{\mathcal{X}}$ -integrable function $h : \mathcal{X} \rightarrow \mathbb{R}^+$ such that $g(x; \theta) \leq h(x)$ over \mathcal{X} for all $\theta \in \Theta_0$.

This assumption is weak, and is satisfied for example as soon as the target density is bounded over its support. It does, nevertheless, exclude some well-known distributions such as, e.g., the arcsine distribution.

3.2. Main result

With these notations, we are ready to state and prove our general characterization theorem.

Theorem 3.1. Let $g \in \mathcal{G}$, let Z_θ be distributed according to $g(\cdot; \theta)$, and let X be a random vector taking values on \mathcal{X} . Fix an interior point $\theta_0 \in \Theta$. Then the following two assertions hold.

- (1) If $X \stackrel{\mathcal{L}}{=} Z_{\theta_0}$, then $\mathbb{E}[\mathcal{T}_{\theta_0}(f, g)(X)] = 0$ for all $f \in \mathcal{F}(g; \theta_0)$.
- (2) If g also satisfies Assumption A at θ_0 and if $\mathbb{E}[\mathcal{T}_{\theta_0}(f, g)(X)] = 0$ for all $f \in \mathcal{F}(g; \theta_0)$, then

$$X \mid X \in S_{\theta_0} \stackrel{\mathcal{L}}{=} Z_{\theta_0}. \quad (3.2)$$

The first statement in Theorem 3.1 is standard; it implies that in order to obtain a Stein operator for a given θ -parametric density g at a point θ_0 , it suffices to find a collection of functions f such that the conditions in Definition 3.1

hold and then apply the operator given in Definition 3.2. As we will show in the next section, this allows for recovering many well-known Stein operators, and for constructing many more. The second statement is also quite standard whenever $S_{\theta_0} = \mathcal{X}$. If $S_{\theta_0} \subset \mathcal{X}$, then things are slightly more tricky. Indeed, in this case, equation (3.2) does not imply that the law of X is necessarily that of Z_{θ_0} , but rather that if the distribution of X has support S_{θ_0} and if X satisfies $\mathbb{E}[\mathcal{T}_{\theta_0}(f, g)(X)] = 0$ (on S_{θ_0} by definition of $\mathcal{T}_{\theta_0}(f, g)$) for all $f \in \mathcal{F}(g; \theta_0)$, then X is distributed according to $g(\cdot; \theta_0)$. This is in accordance with all other results of this form.

Proof. (1) Since Condition (iii) allows for differentiating w.r.t. θ under the integral in Condition (i) and since differentiating w.r.t. θ is allowed thanks to Condition (ii), the claim follows immediately.

(2) First suppose that $p = 1$, and fix $\Theta_0 \subset \Theta$, a bounded (rectangular) neighborhood of θ_0 on which g satisfies Assumption A at θ_0 . Define, for $A \in \mathcal{B}_{\mathcal{X}}$, the mapping

$$f_A : \mathcal{X} \times \Theta_0 \rightarrow \mathbb{R} : (x, \theta) \mapsto \frac{1}{g(x; \theta)} \int_{\theta_0}^{\theta} l_A(x; u, \theta) g(x; u) dm_{\Theta}(u) \quad (3.3)$$

with

$$l_A(x; u, \theta) := (\mathbb{I}_A(x) - \mathbb{P}(Z_u \in A | Z_u \in S_{\theta})) \mathbb{I}_{S_{\theta}}(x),$$

where

$$\mathbb{P}(Z_u \in B) = \int_{\mathcal{X}} \mathbb{I}_B(x) g(x; u) dm_{\mathcal{X}}(x)$$

for $B \in \mathcal{B}_{\mathcal{X}}$. Note that, for the event $[Z_u \in S_{\theta}]$ to have a non-zero probability, it is crucial to work in a neighborhood Θ_0 rather than in Θ ; clearly, this event is always true when S_{θ} does not depend on θ . To see that f_A belongs to $\mathcal{F}(g; \theta_0)$, first note that

$$\begin{aligned} \int_{\mathcal{X}} f_A(x; \theta) g(x; \theta) dm_{\mathcal{X}}(x) &= \int_{\mathcal{X}} \int_{\theta_0}^{\theta} l_A(x; u, \theta) g(x; u) dm_{\Theta}(u) dm_{\mathcal{X}}(x) \\ &= \int_{\theta_0}^{\theta} \int_{\mathcal{X}} l_A(x; u, \theta) g(x; u) dm_{\mathcal{X}}(x) dm_{\Theta}(u), \end{aligned}$$

where the last equality follows from Fubini's theorem, which can be applied for all $\theta \in \Theta_0$, since in this case there exists a constant M such that

$$\int_{\theta_0}^{\theta} \int_{\mathcal{X}} |l_A(x; u, \theta)| g(x; u) dm_{\mathcal{X}}(x) dm_{\Theta}(u) \leq 2|\theta - \theta_0| \leq M$$

for all $\theta \in \Theta_0$. We also have, by definition of l_A ,

$$\begin{aligned} \int_{\mathcal{X}} l_A(x; u, \theta) g(x; u) dm_{\mathcal{X}}(x) &= \mathbb{P}(Z_u \in A \cap S_{\theta}) - \mathbb{P}(Z_u \in A | Z_u \in S_{\theta}) \mathbb{P}(Z_u \in S_{\theta}) \\ &= 0. \end{aligned}$$

Hence f_A satisfies Condition (i). Condition (ii) is easily checked. Regarding Condition (iii), one sees that

$$\partial_t (f_A(x; t)g(x; t))|_{t=\theta} = l_A(x; \theta, \theta)g(x; \theta) + H(x; \theta), \quad (3.4)$$

with $H(x; \theta)$ a function whose complete expression is provided in the Appendix. As shown there, it is easy to bound $H(x; \theta)$ uniformly in θ over Θ_0 by a $m_{\mathcal{X}}$ -integrable function. Moreover, Assumption A guarantees that the same holds for $l_A(x; \theta, \theta)g(x; \theta)$. Hence f_A satisfies Condition (iii). Wrapping up, we have thus proved that $f_A \in \mathcal{F}(g; \theta_0)$. The conclusion follows, since $H(x; \theta_0) = 0$ for all $x \in \mathcal{X}$ (see the Appendix) and since, by hypothesis,

$$\mathbb{E}[\mathcal{T}_{\theta_0}(f_A, g)(X)] = \mathbb{E}[\mathbb{I}_{A \cap S_{\theta_0}}(X) - \mathbb{P}(Z_{\theta_0} \in A) \mathbb{I}_{S_{\theta_0}}(X)] = 0.$$

Next suppose that $p > 1$. Let $\theta_0 := (\theta_0^1, \dots, \theta_0^p)$ and fix $\Theta_0 := \Theta_0^1 \times \dots \times \Theta_0^p$ a bounded (rectangular) neighborhood of θ_0 on which g satisfies Assumption A at θ_0 . Define, for all $j = 1, \dots, p$ and for all $A \in \mathcal{B}_{\mathcal{X}}$, the mappings

$$\bar{\theta}_0^j : \Theta_0^j \rightarrow \Theta_0 : u \mapsto (\theta_0^1, \dots, \theta_0^{j-1}, u, \theta_0^{j+1}, \dots, \theta_0^p)$$

and

$$f_A^j : \mathcal{X} \times \Theta_0 \rightarrow \mathbb{R} : (x, \theta) \mapsto \frac{1}{g(x; \theta)} \int_{\Theta_0^j} l_A^j(x; u, \theta^j) g(x; \bar{\theta}_0^j(u)) dm_{\Theta^j}(u),$$

with

$$l_A^j(x; u, \theta^j) := \left(\mathbb{I}_A(x) - \mathbb{P} \left(Z_u^j \in A \mid Z_u^j \in S_{\bar{\theta}_0^j(\theta^j)} \right) \right) \mathbb{I}_{S_{\bar{\theta}_0^j(\theta^j)}}(x),$$

where

$$\mathbb{P}(Z_u^j \in B) := \int_{\mathcal{X}} \mathbb{I}_B(x) g(x; \bar{\theta}_0^j(u)) dm_{\mathcal{X}}(x)$$

for $B \in \mathcal{B}_{\mathcal{X}}$. The p -variate equivalent of the function f_A in (3.3) is given by $f_A^{(p)}(x; \theta) := \sum_{j=1}^p f_A^j(x; \theta)$. Along the same lines as for the special case $p = 1$, Conditions (i)-(iii) are now easily seen to be satisfied by $f_A^{(p)}$ (we draw the reader's attention to the fact that the rectangular nature of the neighborhood Θ_0 is important in order to ensure Condition (iii)). The result readily follows. \square

Remark 3.1. Nowhere in the proof did we need to specify whether the random vector X is univariate (for $k = 1$) or multivariate (for $k > 1$).

Remark 3.2. When $p > 1$, the (vectorial) operator $\mathcal{T}_{\theta_0}(f, g)$ contains, in a sense, p different characterizations of the $\theta = (\theta^1, \dots, \theta^p)$ -parametric density g at θ_0 . The requirements (in this formulation of the result) on the test functions f are, perhaps, unnecessarily stringent. Indeed, setting $\theta^{(q)} := (\theta^{i_1}, \dots, \theta^{i_q})$ for $1 \leq i_1 \leq \dots \leq i_q \leq p$, we can obviously consider g as a $\theta^{(q)}$ -parametric density. The corresponding q -dimensional sub-vector of $\mathcal{T}_{\theta_0}(f, g)$ also gives rise to a (vectorial) Stein operator for which the conclusions of Theorem 3.1 also hold

at θ_0 , this time with a possibly larger class of test functions f (thanks to the weakening of the requirements imposed by Condition (iii)). In particular, taking $q = 1$, we obtain p distinct one-dimensional characterizations of g at θ_0 . This might be very helpful in approximation theorems concerning g .

Remark 3.3. Note that both implications in Theorem 3.1 are obtained at fixed $\theta_0 \in \Theta$. We attract the reader's attention to the fact that all our calculations and manipulations, as well as all the conditions on the functions at play, are consequently local around θ_0 .

Remark 3.4. All the definitions and arguments above can be extended to encompass distributions with a discrete parameter space Θ (such as, e.g., the discrete uniform). For this it suffices, in a sense, to replace the derivatives and integrals by forward (or backward) differences and summations, respectively. Although it is easy to obtain Stein operators by this means, determining the exact conditions under which the theorem holds nevertheless requires some care, since in this case there arise problems which originate in the interplay between the support of the target density and the parameter of interest. Because of these (structural) intricacies, working out explicit conditions on the target density in this framework appears to be a rather sterile exercise, which is perhaps better suited to ad hoc case by case arguments. This issue will no longer be addressed within the present paper.

The first statement of Theorem 3.1 can be seen as a user-friendly Stein operator-producing mechanism, since any subclass $\tilde{\mathcal{F}}(g; \theta_0) \subset \mathcal{F}(g; \theta_0)$ yields a left-right implication, i.e. an implication of the form

$$X \sim g(\cdot; \theta_0) \implies \mathbb{E}[\mathcal{T}_{\theta_0}(f, g)(X)] = 0 \text{ for all } f \in \tilde{\mathcal{F}}(g; \theta_0).$$

This raises some important questions. Indeed, consider for instance the two operators provided in Example 3.2. As it turns out, both these operators have proven to be extremely useful in applications and their properties are fundamental in the history of the Stein method. However, as already noted by a number of authors before us, they are by no means the only such operators for the Gaussian or the Poisson distribution; in our framework they are just two particular instances of equation (3.1) restricted to certain very specific forms of test functions. A natural question is therefore that of whether there exist other subclasses of test functions for which the corresponding operators would also be useful in applications. It is possible that this question does not allow for a fully satisfactory answer. More precisely it is possible that, for any given problem, there is no *a priori* reason why a given operator would yield better rates of convergence than any other, and perhaps in each problem a careful combination of different characterizations (*à la* Chatterjee, Fulman and Röllin [6]) would be fruitful and would allow for obtaining better results than those obtained by focusing on a single characterization alone.

In any case it seems intuitively clear that, in order for a subclass and the corresponding operator to be of practical use, they need to characterize the law

under consideration, that is, we should have the relationship

$$X \sim g(\cdot; \theta_0) \iff \mathbb{E}[\mathcal{T}_{\theta_0}(f, g)(X)] = 0 \text{ for all } f \in \tilde{\mathcal{F}}(g; \theta_0),$$

where the right-left implication is to be understood in the sense of (3.2) in case S_{θ_0} is a strict subset of \mathcal{X} . Constructing such subclasses, which we call *θ -characterizing for g at θ_0* , is relatively easy. Indeed it suffices to adjoin the function f_A defined in (3.3) to any collection (even empty) of test functions which satisfy the three conditions in Definition 3.1. Such an approach is, however, of limited interest and, moreover, does not allow for clearly identifying the form of the corresponding operators. We therefore suggest a more constructive approach, which we describe in detail in the next section.

4. Characterizing probability distributions

In this section we provide a general “recipe” which allows for constructing θ -characterizing subclasses with well-identified operators. We apply our method to build general characterizations for location families, scale families and discrete distributions. Many well-known Stein characterizations fall under the umbrella of these results. We also show how our method can be applied to obtain more unusual characterizations.

4.1. Characterizations under an exchangeability condition

For the sake of simplicity, we let $k = p = 1$. Fix $\theta_0 \in \Theta$ and choose $g \in \mathcal{G}$ which satisfies Assumption A at θ_0 . In order to construct a θ -characterizing subclass $\tilde{\mathcal{F}}(g; \theta_0) \subset \mathcal{F}(g; \theta_0)$, we suggest the following method.

Step 1: Consider Condition (i) in Definition 3.1, which requires that we have

$$\int_{\mathcal{X}} f(x; \theta) g(x; \theta) dm_{\mathcal{X}}(x) = c_f$$

for $c_f \in \mathbb{R}$. In many cases, the interaction between the variable x and the parameter θ within the density g allows to determine a favored family of test functions $\tilde{f}_0(x; \theta)$ which satisfy this condition. Moreover, these functions are usually expressible as $\tilde{f}_0(x; \theta) = \tilde{T}(f_0; \theta)(x)$, with $f_0 \in \mathcal{X}^*$ and $\tilde{T} : \mathcal{X}^* \times \Theta \rightarrow (\mathcal{X} \times \Theta)^*$.

Step 2: For \tilde{T} and f_0 as given in Step 1, define the *exchanging operator* $T : \mathcal{X}^* \times \Theta \rightarrow (\mathcal{X} \times \Theta)^*$ as a transformation which satisfies the *exchangeability condition*

$$\partial_{\theta} \left(\tilde{T}(f_0; \theta)(x) g(x; \theta) \right) \Big|_{\theta=\theta_0} = \partial_y (T(f_0; \theta_0)(y) g(y; \theta_0)) \Big|_{y=x} \quad (4.1)$$

over \mathcal{X} , where ∂_y either means the derivative in the sense of distributions or the discrete (forward or backward) difference, and we hereby implicitly require that

T is such that the derivative on the rhs of (4.1) is well-defined over \mathcal{X} .

Step 3: Define the class $\mathcal{F}_0 := \mathcal{F}_0(g; \theta_0)$ as the collection of all functions $f_0 \in \mathcal{X}^*$ such that $\tilde{T}(f_0; \theta) \in \mathcal{F}(g; \theta_0)$. Note that we therefore have the (new) left-right implication

$$X \sim g(\cdot; \theta_0) \implies \mathbb{E} \left[\frac{\partial_y (T(f_0; \theta_0)(y) g(y; \theta_0))|_{y=X}}{g(X; \theta_0)} \right] = 0 \text{ for all } f_0 \in \mathcal{F}_0.$$

Step 4: Solve the *Stein equation*

$$\partial_y (T(f_0^A; \theta_0)(y) g(y; \theta_0))|_{y=x} = l_A(x; \theta_0, \theta_0) g(x; \theta_0) \quad (4.2)$$

where $l_A(x; \theta_0, \theta_0)$ is as in the proof of Theorem 3.1. If $T(\cdot; \theta_0)$ is invertible, it then suffices to check whether the corresponding f_0^A belongs to \mathcal{F}_0 in order to obtain the characterization

$$X \sim g(\cdot; \theta_0) \iff \mathbb{E} \left[\frac{\partial_y (T(f_0; \theta_0)(y) g(y; \theta_0))|_{y=X}}{g(X; \theta_0)} \right] = 0 \text{ for all } f_0 \in \mathcal{F}_0,$$

where the right-left implication is to be understood, as before, in the sense of (3.2) in case the support S_{θ_0} of $g(\cdot; \theta_0)$ is a strict subset of \mathcal{X} .

The resulting θ -characterizing subclass $\tilde{\mathcal{F}}(g; \theta_0)$ is none other than the collection $\{\tilde{T}(f_0; \theta) \mid f_0 \in \mathcal{F}_0\} \cup \{f_A\}$; this collection not only has the desired properties, but also is accompanied with a well-identified Stein operator. In the sequel it will be more convenient to state our results in terms of \mathcal{F}_0 rather than in terms of $\tilde{\mathcal{F}}(g; \theta_0)$. This is in accordance with all other results of this form.

There are a number of ways in which one can extend the method presented above to the cases $k > 1$ and $p > 1$. Also, for given θ_0 and θ -parametric density g , the choice of class \mathcal{F}_0 and exchanging operator T is not unique. Moreover, determining straightforward minimal conditions on the f_0 for the characterization to hold seems to be impossible without making further regularity assumptions on the target density g . These considerations entail that it is perhaps more fruitful to tackle different θ -parametric densities with ad hoc arguments. There are nevertheless important instances in which one can obtain general results with relative ease. To this end consider the following assumption on univariate θ -parametric densities.

Assumption B : there exists $x_0 \in \mathcal{X}$ such that

$$\left| \int_{\mathcal{X}} \left(\int_{x_0}^x l_A(y; \theta_0, \theta_0) g(y; \theta_0) dm_{\mathcal{X}}(y) \right) \mathbb{I}_{S_{\theta_0}}(x) dm_{\mathcal{X}}(x) \right| < \infty$$

for all $A \in \mathcal{B}_{\mathcal{X}}$, where $l_A(y; \theta_0, \theta_0)$ is defined as in the proof of Theorem 3.1.

This is a condition on the tails of the density $g(\cdot; \theta_0)$ which is, for instance, satisfied by the Gaussian and the exponential distributions (while the latter is evident, see for the former [10] page 4). As we will see, Assumption B is useful for determining general characterization results in location and scale models.

4.2. Location-based characterizations

In this subsection we apply the method described in Section 4.1 to study laws whose parameter of interest is a location parameter.

Corollary 4.1. *Let $k = p = 1$ and $\mathcal{X} = \mathbb{R} = \Theta$, and fix $\mu_0 \in \Theta$. Define \mathcal{G}_{loc} as the collection of densities $g_0 : \mathcal{X} \rightarrow \mathbb{R}^+$ with support $S \subset \mathcal{X}$ such that the μ -parametric density $g(x; \mu) = g_0(x - \mu)$ belongs to \mathcal{G} and satisfies Assumptions A and B at μ_0 . Let $\Theta_0 \subset \Theta$ be as in Assumption A, and define $\mathcal{F}_0 := \mathcal{F}_0(g_0; \mu_0)$ as the collection of all $f_0 : \mathcal{X} \rightarrow \mathbb{R}$ such that*

Condition (μ -i) : $|\int_{\mathcal{X}} f_0(x) g_0(x) dm_{\mathcal{X}}(x)| < \infty$,

Condition (μ -ii) : the mapping $x \mapsto f_0(x) g_0(x)$ is differentiable in the sense of distributions over \mathcal{X} ,

Condition (μ -iii) : there exists a $m_{\mathcal{X}}$ -integrable function $h : \mathcal{X} \rightarrow \mathbb{R}^+$ such that $|\partial_y(f_0(y - \mu) g_0(y - \mu))|_{y=x} \leq h(x)$ over \mathcal{X} for all $\mu \in \Theta_0$.

Then \mathcal{F}_0 is μ -characterizing for g_0 at μ_0 , with μ -characterizing operator

$$\mathcal{T}_{\mu_0}(f_0, g_0) : \mathcal{X} \rightarrow \mathcal{X} : x \mapsto -\frac{\partial_y(f_0(y - \mu_0) g_0(y - \mu_0))|_{y=x}}{g_0(x - \mu_0)}. \quad (4.3)$$

A proof, which is a direct application of the method described in Section 4.1, is provided in the Appendix. The operator in (4.3) – as well as the conditions on the densities and the conditions on the test functions f_0 – differ slightly from those already available in the literature; this matter has already been discussed in Section 2.

Corollary 4.1 contains a number of well-known univariate characterizations covered in the literature. For instance, taking $g(\cdot; \mu)$ to be the density of a $\mathcal{N}(\mu, 1)$ (which satisfies Assumptions A and B at $\mu_0 = 0$) we can use the operator provided in Example 3.2; Corollary 4.1 then leads to the famous Stein characterization of the standard normal distribution. Likewise, introducing an artificial location parameter μ within the exponential density with scale parameter 1 (which, again, satisfies Assumptions A and B at $\mu_0 = 0$) leads to the characterization of the exponential distribution given in Example 2.2. More generally, when g belongs to the (continuous) *exponential family* (see [19]), one easily sees how the same manipulations allow to retrieve the known characterizations (see also [20] or [24]). We refer to [35], [12] and [32] for more location-based characterizations.

Next consider the semi-circular law whose density is given by

$$g_0(x - \mu) = \frac{2}{\pi \sigma^2} \sqrt{\sigma^2 - (x - \mu)^2} \mathbb{I}_{[-\sigma, \sigma]}(x - \mu), \quad (4.4)$$

with $\mu \in \mathbb{R}$ being a location and $\sigma \in \mathbb{R}_0^+$ a known scale parameter. In the special case $\mu = 0$ and $\sigma = 2$, Götze and Tikhomirov [17] prove that a random variable X is distributed according to (4.4) if and only if

$$\mathbb{E}[(4 - X^2)f'(X) - 3Xf(X)] = 0 \quad (4.5)$$

for all test functions f in a certain class of functions. We claim that (4.5) falls within the category of location-based characterizations. To see this it suffices to note that, although we are in a location model with target density satisfying Assumptions A and B at all points $\mu_0 \in \mathbb{R}$, the derivative $g'_0(x - \mu)$ is not bounded at the edges of the support. Conditions (μ -ii) and (μ -iii) therefore entail some stringent requirements on the admissible class of test functions. In order to be able to read these requirements more easily, one way to proceed is to consider only f_0 's of the form $f_0(x) = f_1(x)(\sigma^2 - x^2)^r$, with $r > 1/2$. Writing out the location-based characterization in terms of the functions f_1 instead of f_0 yields, for $r = 1$, the expression in (4.5); sufficient conditions on f_1 for f_0 to belong to \mathcal{F}_0 are easy to provide (see [17] in the case $r = 1$ and $\sigma = 2$).

Note that, when the target density belongs to Pearson's family of distributions, there exists a general result due to [32] for obtaining Stein characterizations which encompasses many of the characterizations obtainable through Stein's density approach. We wish to stress the fact that all these results can be recovered through our Corollary 4.1.

And now a multivariate example. Consider a random k -vector Z_{μ_0} with $\mu_0 \in \mathbb{R}^k$ and density of the form $g(x; \mu) := g_0(x - \mu) = g_0(x_1 - \mu^1, x_2 - \mu^2, \dots, x_k - \mu^k)$. Suppose, for the sake of simplicity, that the support of $g_0(x - \mu)$ does not depend on μ (i.e. $S = \mathcal{X} = \mathbb{R}^k$). One way to characterize such distributions at μ_0 is to define, for fixed x_2, \dots, x_k , the univariate μ^1 -parametric density $g_1(x_1; \mu_1) = g_0(x_1 - \mu^1, x_2 - \mu_0^2, \dots, x_k - \mu_0^k)$. Requiring that $g_1 \in \mathcal{G}$ and satisfies Assumptions A and B at μ_0^1 , we easily determine a class of functions \mathcal{F}_0^1 as in Corollary 4.1 to obtain

$$\begin{aligned} X &\stackrel{\mathcal{L}}{=} Z_{\mu_0} \iff \\ \mathbb{E} \left[\frac{\partial_y (f_0(y - \mu_0, X) g_0(y - \mu_0, X))|_{y=X_1}}{g_0(X - \mu_0)} \right] &= 0 \text{ for all } f_0 \in \mathcal{F}_0^1, \end{aligned} \quad (4.6)$$

where we use the abuse of notations $(y - \mu_0, X) = (y - \mu_0^1, X_2 - \mu_0^2, \dots, X_k - \mu_0^k)$ and $X - \mu_0 = (X_1 - \mu_0^1, X_2 - \mu_0^2, \dots, X_k - \mu_0^k)$. The choice of μ^1 as parameter of interest was of course for convenience only, and similar relationships hold for derivatives with respect to x_2, \dots, x_k as well. Moreover, when Z_{θ_0} has support \mathcal{X} and independent marginals, one easily sees how to aggregate these different results and write out a class of functions $\mathcal{F}_0^{(k)}$ as in Corollary 4.1 to get

$$X \sim g(\cdot; \theta_0) \iff \mathbb{E} \left[\frac{\nabla_y (f_0(y - \mu_0) g_0(y - \mu_0))|_{y=X}}{g_0(X - \mu_0)} \right] = 0 \text{ for all } f_0 \in \mathcal{F}_0^{(k)}. \quad (4.7)$$

We conclude this section by showing how (4.6) and (4.7) read in the Gaussian case. Here, setting $\mu_0 = 0 \in \mathbb{R}^k$ and plugging the multivariate Gaussian density $g(x; \mu, \Sigma)$ with Σ a known symmetric positive definite $k \times k$ matrix into (4.6) we get, for $j = 1, \dots, k$,

$$X \sim \mathcal{N}(0, \Sigma) \iff \mathbb{E} \left[\partial_{y_j}(f_0(y_j, X)) \Big|_{y_j=X_j} - \sigma_j f_0(X) \right] = 0 \text{ for all } f_0 \in \mathcal{F}_0^j \quad (4.8)$$

where we use the notations $(y_j, X) = (X_1, \dots, X_{j-1}, y_j, X_{j+1}, \dots, X_k)$ and $\sigma_j := (\Sigma^{-1}X)_j = \sum_{i=1}^k (\Sigma^{-1})_{ji} X_i$. Moreover, when Σ is the identity matrix I_k we can use (4.7) to obtain

$$X \sim \mathcal{N}(0, I_k) \iff \mathbb{E} \left[\nabla_y(f_0(y)) \Big|_{y=X} - X f_0(X) \right] = 0 \text{ for all } f_0 \in \mathcal{F}_0^{(k)}. \quad (4.9)$$

These characterizations of the multivariate Gaussian are, to the best of our knowledge, new. They are to be compared with existing results given, e.g., in [7] and [30].

4.3. Scale-based characterizations

In this subsection we apply the method described in Section 4.1 to study laws whose parameter of interest is a scale parameter.

Corollary 4.2. *Let $k = p = 1$, $\mathcal{X} = \mathbb{R}$ and $\Theta = \mathbb{R}_0^+$, and fix $\sigma_0 \in \Theta$. Define \mathcal{G}_{sca} as the collection of densities $g_0 : \mathcal{X} \rightarrow \mathbb{R}^+$ with support $S \subset \mathcal{X}$ such that the σ -parametric density $g(x; \sigma) = \sigma g_0(\sigma x)$ belongs to \mathcal{G} and satisfies Assumptions A and B at σ_0 . Let $\Theta_0 \subset \Theta$ be as in Assumption A, and define $\mathcal{F}_0 := \mathcal{F}_0(g_0; \sigma_0)$ as the collection of all $f_0 : \mathcal{X} \rightarrow \mathbb{R}$ such that*

Condition (σ -i) : $\left| \int_{\mathcal{X}} f_0(x) g_0(x) dm_{\mathcal{X}}(x) \right| < \infty$.

Condition (σ -ii) : the mapping $x \mapsto x f_0(x) g_0(x)$ is differentiable in the sense of distributions over \mathcal{X} ,

Condition (σ -iii) : there exists a $m_{\mathcal{X}}$ -integrable function $h : \mathcal{X} \rightarrow \mathbb{R}^+$ such that $\left| \partial_y(y f_0(\sigma y) g_0(\sigma y)) \Big|_{y=x} \right| \leq h(x)$ over \mathcal{X} for all $\sigma \in \Theta_0$.

Then \mathcal{F}_0 is σ -characterizing for g_0 at σ_0 , with σ -characterizing operator

$$\mathcal{T}_{\sigma_0}(f_0, g_0) : \mathcal{X} \rightarrow \mathcal{X} : x \mapsto \frac{\partial_y(y f_0(\sigma_0 y) g_0(\sigma_0 y)) \Big|_{y=x}}{\sigma_0 g_0(\sigma_0 x)}. \quad (4.10)$$

The proof of Corollary 4.2 is similar to that of Corollary 4.1, and hence is omitted.

As in the location case, this result can be extended in a number of ways to the multivariate setting. In the univariate setup, if g is the exponential density with scale parameter λ and if λ_0 is set to 1, we retrieve the characterization (2.5). If g_0 is the density of a $\mathcal{N}(0, 1)$ distribution, the above characterization reads

$$X \sim \mathcal{N}(0, 1) \iff \mathbb{E}[X f'_0(X) + (1 - X^2) f_0(X)] = 0 \quad (4.11)$$

for all (differentiable) $f_0 \in \mathcal{F}_0$.

4.4. Discrete characterizations

Our last general result concerns discrete distributions. In this instance there is, in general, no unique interpretation of the parameters of interest; it depends on the law under investigation. As will be clear from the proof of Corollary 4.3 below (see the Appendix), our approach in this setting allows us to dispense with Assumption B, which was needed in order to ensure Condition (i) in Definition 3.1. However we need to strengthen Assumption A as follows.

Assumption A' : for $\psi(x; \theta) := \partial_u (g(x; u)/g(0; u))|_{u=\theta}$, there exists a neighborhood Θ_0 of θ_0 and a summable function $h : \mathbb{Z} \rightarrow \mathbb{R}^+$ such that

$$\left| \Delta_x^+ \left(\frac{\psi(x; \theta)}{\psi(x; \theta_0)} \sum_{j=0}^{x-1} l_A(j; \theta_0, \theta_0) g(j; \theta_0) \right) \right| \leq h(x)$$

over \mathcal{X} for all $\theta \in \Theta_0$ and for all $A \in \mathcal{B}_{\mathcal{X}}$, where $l_A(j; \theta_0, \theta_0)$ is defined as in the proof of Theorem 3.1 and where Δ_x^+ is the forward difference with respect to x .

Assumption A' is sufficient to ensure Condition (iii) in the discrete setting. It is not restrictive and is satisfied by all the (discrete) distributions we have considered. For example, in the Poisson case, the ratio $\psi(x; \theta)/\psi(x; \theta_0)$ is none other than $(\lambda/\lambda_0)^{x-1} \mathbb{I}_{\mathbb{N}_0}(x)$ so that known arguments (see page 65 of [15]) apply.

Corollary 4.3. *Let $k = p = 1$, $\mathcal{X} = \mathbb{Z}$ and $\Theta \subset \mathbb{R}$, and fix $\theta_0 \in \Theta$. Define \mathcal{G}_{dis} as the collection of θ -parametric discrete densities $g(\cdot; \theta) : \mathcal{X} \rightarrow [0, 1]$ with support $S \subset \mathcal{X}$, which we take of the form $S = [N] := \{0, \dots, N\}$ for some $N \in \mathbb{N}_0 \cup \{\infty\}$ not depending on θ , such that $g \in \mathcal{G}$ and satisfies Assumption A' at θ_0 . Define \mathcal{F}_0 as the collection of all functions $f_0 : \mathcal{X} \rightarrow \mathbb{R}$ for which there exists a summable function $h : \mathbb{Z} \rightarrow \mathbb{R}^+$ such that $|\Delta_x^+(f_0(x) \partial_u (g(x; u)/g(0; u))|_{u=\theta})| \leq h(x)$ over \mathcal{X} for all $\theta \in \Theta_0$, with Θ_0 as in Assumption A'.*

Then \mathcal{F}_0 is θ -characterizing for g at θ_0 , with θ -characterizing operator

$$\mathcal{T}_{\theta_0}(f_0, g)(x) = \frac{\Delta_x^+ \left(f_0(x) \partial_\theta (g(x; \theta)/g(0; \theta))|_{\theta=\theta_0} \right)}{g(x; \theta_0)}.$$

Corollary 4.3 contains a number of well-known discrete characterizations covered in the literature among which, for instance, those for the Poisson (see the operator in Example 3.2), the geometric $Geom(p)$, with p -characterizing operator

$$\mathcal{T}_p(f_0, g)(x) = -\frac{1}{p} \left((x+1)f_0(x+1) - \frac{x}{1-p}f_0(x) \right) \mathbb{I}_{\mathbb{N}}(x),$$

or the binomial $Bin(n, p)$, with p -characterizing operator

$$\mathcal{T}_p(f_0, g)(x) = (1-p)^{-n-2} \left((n-x)f_0(x+1) - \frac{1-p}{p}xf_0(x) \right) \mathbb{I}_{[n]}(x).$$

The same arguments allow, of course, for dealing with other perhaps more exotic discrete distributions. Consider, for the sake of illustration, the case of the multinomial $M(n, p_1, \dots, p_k)$, with density

$$g(x) = \frac{n!}{\prod_{j=0}^k x_j!} \prod_{j=0}^k p_j^{x_j} \mathbb{I}_{\Delta^n}(x) \quad (4.12)$$

where $x_0 = n - \sum_{j=1}^k x_j$, $p_0 = 1 - \sum_{j=1}^k p_j$ and

$$\Delta^n = \{(x_1, \dots, x_k) \in \mathbb{N}^k \mid 0 \leq x_1 + \dots + x_k \leq n\}.$$

In the same spirit as our previous multivariate characterizations, we start by transforming the problem into a univariate one. For this choose p_1 to be the parameter of interest, and rewrite (4.12) as

$$g(x) = \left(\binom{\bar{n}_1}{x_1} p_1^{x_1} (\bar{p}_1 - p_1)^{\bar{n}_1 - x_1} \right) \frac{n!}{\bar{n}_1!} \frac{\prod_{j=2}^k p_j^{x_j}}{\prod_{j=2}^k x_j!} \mathbb{I}_{\Delta^n}(x)$$

where, letting $\bar{x}_1 = \sum_{j=2}^k x_j$, we denote $\bar{n}_1 = n - \bar{x}_1$ and $\bar{p}_1 = 1 - \sum_{j=2}^k p_j$. Straightforward computations readily yield the corresponding operator

$$\mathcal{T}_{p_1}(f_0, g)(x) = \xi(x; n) \left((\bar{n}_1 - x_1) f_0(x_1 + 1) - \frac{\bar{p}_1 - p_1}{p_1} x_1 f_0(x_1) \right) \mathbb{I}_{\Delta^n}(x),$$

with

$$\xi(x; n) = \frac{\bar{p}_1}{(\bar{p}_1 - p_1)^{\bar{n}_1 + 2}}.$$

In each of the above cases, determining sufficient conditions on the test functions f_0 for the operators to be θ -characterizing is now a simple exercise which is left to the reader.

5. Uncovering new results

In this final section, we tackle two examples which do not fall within the scope of the previous general results. In each case, we try to convey some intuition as to how our method works. As will appear, each of these cases requires the development of *ad hoc* arguments.

5.1. The uniform distribution

First take the target distribution g to be the density of a uniform $U[a, b]$ for $a \leq b \in \mathbb{R}$, and define a to be the parameter of interest. This law is not, *stricto sensu*, a member of the scale family. It is, however, easily seen that it belongs to \mathcal{G} for all $a \neq b$ and satisfies Assumptions A and B at all $a < b$, with b fixed. It is readily seen that the exchanging operator $T(f_0; a)(x) = (x - b)/(b - a) f_0((x - a)/(b - a))$ yields the precious relationship (4.1), with $\bar{T}(f_0; a) = f_0((x - a)/(b - a))$. This leads to the following result (the proof is left to the reader).

Corollary 5.1. *Let \mathcal{F}_0 be the collection of all functions $f_0 : \mathbb{R} \rightarrow \mathbb{R}$ which are differentiable (in the sense of distributions) on $[0, 1]$. Then \mathcal{F}_0 is a -characterizing for g , with a -characterizing operator*

$$\mathcal{T}_a(f, g)(x) = \frac{1}{b-a} \left(\frac{x-b}{b-a} f'_0 \left(\frac{x-a}{b-a} \right) + f_0 \left(\frac{x-a}{b-a} \right) \right) \mathbb{I}_{[a,b]}(x) - f_0(0)$$

for all $f_0 \in \mathcal{F}_0$.

Similarly, one can also construct a b -characterizing operator and a b -characterization for the uniform law on $[a, b]$. A third way to characterize this law is to proceed as in Section 4.2 and construct a μ -characterization, for μ a location parameter introduced by considering the density $g(x - \mu)$ and working, through Corollary 4.1, with respect to μ . This yields the expression in Example 2.1.

5.2. The Student distribution

Take the target distribution g to be the density of a Student $T(\nu)$ with parameter of interest $\nu \in \mathbb{R}_0^+$, the tail weight parameter. This law belongs to \mathcal{G} for all $\nu > 0$ and satisfies Assumption A at all $\nu > 0$. It is readily seen that the exchanging operator

$$T(f_0; \nu)(x) = -\frac{1}{2\nu} \frac{\Gamma(\nu/2)}{\Gamma((\nu+1)/2)} x \left(1 + \frac{x^2}{\nu} \right)^{\nu/2} f_0 \left(\frac{x^2}{\nu} \right)$$

yields the precious relationship (4.1), with

$$\tilde{T}(f_0; \nu)(x) = \frac{\Gamma(\nu/2)}{\Gamma((\nu+1)/2)} (1 + x^2/\nu)^{\nu/2} f_0(x^2/\nu).$$

Sufficient conditions on f_0 for the now usual requirements to be fulfilled are easily imposed. This leads to the following result.

Corollary 5.2. *Fix $\nu > 2$. Let \mathcal{F}_0 be the collection of differentiable (in the sense of distributions) functions $f_0 : \mathbb{R} \rightarrow \mathbb{R}$ such that $|f_0(x^2)|/\sqrt{1+x^2}$ and $|xf'_0(x^2)|$ are $m_{\mathbb{R}}$ -integrable. Then \mathcal{F}_0 is ν -characterizing for g , with ν -characterizing operator*

$$\mathcal{T}_\nu(f_0, g)(x) = \xi(x; \nu) \left(2x^2 f'_0 \left(\frac{x^2}{\nu} \right) - f_0 \left(\frac{x^2}{\nu} \right) \left(\frac{x^2}{1 + \frac{x^2}{\nu}} - \nu \right) \right),$$

where $\xi(x; \nu) = -\Gamma(\nu/2)(2\nu^2\Gamma((\nu+1)/2))^{-1} (1 + x^2/\nu)^{\nu/2}$.

The proof of this result is mainly computational and follows along the same lines as that of all other similar results provided in this paper.

It seems appropriate to conclude on this final example. Obviously, similar parameter-based characterizations can be obtained, by means of the same tools, for gamma, hypergeometric, Laplace, Pareto distributions, etc. As far as we know there exists no univariate characterization which cannot be obtained through our approach.

6. Applications

In all works related with Stein's method the characterization is merely the first step in a complicated and not a little mysterious process. In this paper we do not discuss the intricacies and subtleties of the method, and rather refer the non-initiated reader to the monographs [2, 3] or [11] for an overview. Moreover our parametric approach to the characterizations has to this date never been used for any application. The purpose of this section is to provide two simple and direct consequences of our vision. Deeper results are still under investigation.

6.1. Solving Stein equations

Suppose that, for a given parametric target distribution g , we dispose of characterizations of the form $Z \sim g(\cdot; \theta_0) \iff \mathbb{E}[\mathcal{T}_{\theta_0}(f, g)(Z)] = 0$ for all f in $\mathcal{F}(g; \theta_0)$, where $\mathcal{T}_{\theta_0}(f, g)$ is a Stein operators. Then a *Stein equation* for g at θ_0 is a differential equation given by

$$\mathcal{T}_{\theta_0}(f_h, g)(x) = l(x) \quad (6.1)$$

for $l : \mathcal{X} \rightarrow \mathbb{R}$ some function.

Example 6.1. *In the Gaussian case, we obtain the location equation*

$$f'(x) - xf(x) = l(x)$$

and the scale equation

$$xf'(x) + (1 - x^2)f(x) = l(x).$$

In the Exponential case we obtain the location equation

$$(f'(x) - f(x)) \mathbb{I}_{\mathbb{R}^+}(x) = l(x)$$

and the scale equation

$$(xf'(x) - (x - 1)f(x)) \mathbb{I}_{\mathbb{R}^+}(x) = l(x).$$

In the Poisson case we obtain the λ -equation

$$\left(f_0(x + 1) - \frac{x}{\lambda_0} f_0(x) \right) \mathbb{I}_{\mathbb{N}}(x) = l(x).$$

A careful reading of the different proofs provided in this paper shows that Theorem 3.1 not only yields Stein operators, but also solutions to the corresponding Stein equations (see equations (3.3), (A.1) and (A.4)). More specifically, our way of writing the operator (as a single differential) obviously allows for solving *all* such equations in a unified way by simple integration. Note in particular how, in the discrete case, the solution is obtained through straightforward summation. In other words our approach allows for solving *all* Stein equations in a routine fashion.

6.2. Stein's method and information theory

Although there are many consequences to our Theorem 3.1, perhaps the most intuitive is that it provides a hitherto unsuspected direct link between Stein's method and information theoretic tools. Such results are, however, outside the scope and purpose of the present work and will be the subject of separate publications. We nevertheless wish to suggest the flavor of this connection, and therefore conclude the paper with a particularly appealing result.

Choose two parametric densities $p, q \in \mathcal{G}$ sharing the same support S_θ . Take $f \in \mathcal{F}(p; \theta_0)$. We obviously have

$$\begin{aligned} \mathcal{T}_{\theta_0}(f, p)(x) &= \frac{\partial_\theta f(x; \theta) p(x; \theta)|_{\theta=\theta_0}}{p(x; \theta_0)} \\ &= \frac{\partial_\theta f(x; \theta) q(x; \theta)|_{\theta=\theta_0}}{q(x; \theta_0)} \frac{p(x; \theta_0)}{p(x; \theta_0)} + \frac{f(x; \theta_0) q(x; \theta_0)}{p(x; \theta_0)} \partial_\theta \left(\frac{p(x; \theta)}{q(x; \theta)} \right) \Big|_{\theta=\theta_0}. \end{aligned}$$

Straightforward simplifications then yield our final lemma.

Lemma 6.1 (Factorization of Stein operators). *For all $f \in \mathcal{F}(p; \theta_0)$, we have*

$$\mathcal{T}_{\theta_0}(f, p)(x) = \mathcal{T}_{\theta_0}(f, q)(x) + f(x; \theta_0) r_{\theta_0}(p, q)(x), \quad (6.2)$$

with

$$r_{\theta_0}(p, q)(x) := \frac{\partial_\theta p(x; \theta)|_{\theta=\theta_0}}{p(x; \theta_0)} - \frac{\partial_\theta q(x; \theta)|_{\theta=\theta_0}}{q(x; \theta_0)}. \quad (6.3)$$

We call the operator r_{θ_0} a *generalized (standardized) score function* because specifying the role of θ (location, scale, ...) as well as its nature (discrete, continuous) allows to recover a whole family of score functions discussed in [21], [23] or [5]. Such an observation obviously has an intriguing number of immediate applications, but also opens new lines of research which are currently under investigation. See [25] for first results in this direction.

Appendix A: Technical proofs

Proof of equality (3.4). First note that

$$\begin{aligned} &\partial_t (f_A(x; t) g(x; t))|_{t=\theta} \\ &= \partial_t \left(\int_{\theta_0}^t l_A(x; u, t) g(x; u) dm_\Theta(u) \right) \Big|_{t=\theta} \\ &= l_A(x; \theta, \theta) g(x; \theta) + \int_{\theta_0}^\theta \partial_t (l_A(x; u, t))|_{t=\theta} g(x; u) dm_\Theta(u). \end{aligned}$$

Now we have

$$\begin{aligned} \partial_t (l_A(x; u, t))|_{t=\theta} &= \partial_t (\mathbb{I}_{S_t}(x))|_{t=\theta} (\mathbb{I}_A(x) - \mathbb{P}(Z_u \in A | Z_u \in S_\theta)) \\ &\quad - \partial_t (\mathbb{P}(Z_u \in A | Z_u \in S_t))|_{t=\theta} \mathbb{I}_{S_\theta}(x). \end{aligned}$$

On the one hand, we easily see that the function

$$H_1(x; \theta) := \partial_t (\mathbb{I}_{S_t}(x))|_{t=\theta} \int_{\theta_0}^{\theta} (\mathbb{I}_A(x) - \mathbb{P}(Z_u \in A | Z_u \in S_\theta)) g(x; u) dm_\Theta(u)$$

is well-defined, bounded uniformly in θ over Θ_0 by a $m_{\mathcal{X}}$ -integrable function and satisfies $H_1(x; \theta_0) = 0$. On the other hand, we have

$$\begin{aligned} & \partial_t (\mathbb{P}(Z_u \in A | Z_u \in S_t))|_{t=\theta} \\ &= \frac{\partial_t (\mathbb{P}(Z_u \in A \cap S_t))|_{t=\theta}}{\mathbb{P}(Z_u \in S_\theta)} - \partial_t (\mathbb{P}(Z_u \in S_t))|_{t=\theta} \frac{\mathbb{P}(Z_u \in A \cap S_\theta)}{\mathbb{P}(Z_u \in S_\theta)^2}, \end{aligned}$$

where clearly both derivatives are well-defined. Hence the function

$$H_2(x; \theta) := \mathbb{I}_{S_\theta}(x) \int_{\theta_0}^{\theta} \partial_t (\mathbb{P}(Z_u \in A | Z_u \in S_t))|_{t=\theta} g(x; u) dm_\Theta(u)$$

is also well-defined, bounded uniformly in θ over Θ_0 by a $m_{\mathcal{X}}$ -integrable function and satisfies $H_2(x; \theta_0) = 0$. Defining

$$H(x; \theta) := H_1(x; \theta) - H_2(x; \theta)$$

we see that all the assertions in the proof of Theorem 3.1 hold, and, moreover, that

$$\begin{aligned} \partial_t (f_A(x; t)g(x; t))|_{t=\theta_0} &= l_A(x; \theta_0, \theta_0)g(x; \theta_0) + H(x; \theta_0) \\ &= l_A(x; \theta_0, \theta_0)g(x; \theta_0). \end{aligned}$$

This completes the proof of Theorem 3.1. \square

Proof of Corollary 4.1 (location). We apply the method described in Section 4.1.

Step 1: Choose $\tilde{T}(f_0; \mu)(x) = f_0(x - \mu)$.

Step 2: Set $T(f_0; \mu)(x) = -f_0(x - \mu)$.

Step 3: One easily sees that, for any $f_0 \in \mathcal{F}_0$, Conditions $(\mu\text{-i})$ - $(\mu\text{-iii})$ on f_0 entail that Conditions (i)-(iii) are satisfied by $\tilde{T}(f_0; \mu)(x)$.

Step 4: Consider the solution of the Stein equation given by

$$f_0^A(x - \mu_0) = -\frac{1}{g_0(x - \mu_0)} \left(\int_{x_0}^x l_A(y; \mu_0, \mu_0) g_0(y - \mu_0) dm_{\mathcal{X}}(y) + c(x) \right)$$

for some $x_0 \in \mathcal{X}$, where the function $x \mapsto c(x)$ has derivative (in the sense of distributions) equal to zero and is defined in such a way that $(\int_{x_0}^x l_A(y; \mu_0, \mu_0) g_0(y -$

$\mu_0)dm_{\mathcal{X}}(y) + c(x))\partial_x \mathbb{I}_S(x - \mu_0) = 0$ over \mathcal{X} . This function can be expressed as a sum of Dirac delta functions whose vertices are determined by $\partial_x \mathbb{I}_S(x - \mu_0)$. This yields the candidate solution

$$f_0^A(x) = -\frac{1}{g_0(x)} \left(\int_{x_0}^{x+\mu_0} l_A(y; \mu_0, \mu_0) g_0(y - \mu_0) dm_{\mathcal{X}}(y) + c(x + \mu_0) \right). \quad (\text{A.1})$$

For this function to belong to \mathcal{F}_0 , we need Condition $(\mu\text{-ii})$, which is obvious, Condition $(\mu\text{-iii})$, which is also obvious thanks to Assumption A once again, and Condition $(\mu\text{-i})$ which will hold as soon as

$$\begin{aligned} & \left| \int_{\mathcal{X}} f_0^A(x) g_0(x) dm_{\mathcal{X}}(x) \right| \\ &= C + \left| \int_{\mathcal{X}} \int_{x_0}^{x+\mu_0} l_A(y; \mu_0, \mu_0) g(y - \mu_0) dm_{\mathcal{X}}(y) \mathbb{I}_S(x) dm_{\mathcal{X}}(x) \right| < \infty, \end{aligned}$$

where $C = \int_{\mathcal{X}} c(x + \mu_0) \mathbb{I}_S(x) dm_{\mathcal{X}}(x)$ is finite. Since Assumption B then ensures that the quantity $\left| \int_{\mathcal{X}} f_0^A(x) g_0(x) dm_{\mathcal{X}}(x) \right|$ is bounded, Condition $(\mu\text{-i})$ is satisfied as well, which concludes the proof. \square

Proof of Corollary 4.3 (discrete). In this framework, the exchangeability condition (4.1) reads

$$\partial_{\theta}(\tilde{T}(f_0; \theta)(x) g(x; \theta))|_{\theta=\theta_0} = \Delta_x^+(T(f_0; \theta_0)(x) g(x; \theta_0)), \quad (\text{A.2})$$

for some $f_0 \in \mathcal{F}_0$. In order to obtain the announced θ -characterizing operator $\mathcal{T}_{\theta_0}(f_0, g)$, we define

$$\tilde{T}(f_0; \theta)(x) = \frac{\Delta_x^+(f_0(x) g(x; \theta))}{g(x; \theta) g(0; \theta)} \quad (\text{A.3})$$

and the (invertible) exchanging operator

$$T(f_0; \theta_0)(x) = f_0(x) \frac{\partial_{\theta}(g(x; \theta)/g(0; \theta))|_{\theta=\theta_0}}{g(x; \theta_0)}.$$

One readily checks that these choices satisfy the exchangeability condition (A.2).

Fix $\theta_0 \in \Theta$. The sufficient condition is immediate. For the necessary condition to hold, we solve

$$\Delta_x^+(T(f_0^A; \theta_0)(x) g(x; \theta_0)) = l_A(x; \theta_0, \theta_0) g(x; \theta_0),$$

with l_A as before, to obtain the candidate solution

$$f_0^A(x) = (\psi(x; \theta_0))^{-1} \sum_{j=0}^{x-1} l_A(j; \theta_0, \theta_0) g(j; \theta_0), \quad (\text{A.4})$$

where the sum over an empty set is 0. Assumption A' guarantees that this function belongs to \mathcal{F}_0 . \square

Acknowledgements

Christophe Ley's research is supported by a Mandat de Chargé de recherche from the Fonds National de la Recherche Scientifique, Communauté française de Belgique. Christophe Ley is also member of E.C.A.R.E.S. Yvik Swan's research is supported by a Mandat de Chargé de recherche from the Fonds National de la Recherche Scientifique, Communauté française de Belgique.

References

- [1] BARBOUR, A. D. (1988). Stein's method and Poisson process convergence. *J. Appl. Probab.* **25A** 175–184.
- [2] BARBOUR, A. D. AND CHEN, L. H. Y. (2005). *An introduction to Stein's method*. Lecture Notes Series **4**, Institute for Mathematical Sciences, National University of Singapore, Singapore University Press, Singapore.
- [3] BARBOUR, A. D. AND CHEN, L. H. Y. (2005). *Stein's method and applications*. Lecture Notes Series **5**, Institute for Mathematical Sciences, National University of Singapore, Singapore University Press, Singapore.
- [4] BARBOUR, A. D., HOLST, L. AND JANSON, S. (1992). *Poisson approximation*. Oxford Studies in Probability **2**. The Clarendon Press Oxford University Press, New York.
- [5] Barbour, A. D., Johnson, O., Kontoyiannis, I. and Madiman, M. (2010). Compound Poisson approximation via information functionals. *Electronic Journal of Probability* **15** 1344–1369.
- [6] CHATTERJEE, S., FULMAN, J. AND RÖLLIN, A. (2011). Exponential approximation by Stein's method and spectral graph theory. *ALEA Lat. Am. J. Probab. Math. Stat.* To appear.
- [7] CHATTERJEE, S. AND MECKES, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 257–283.
- [8] CHATTERJEE, S. AND SHAO, Q.-M. (2011). Non-normal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *Ann. App. Probab.* **21** 464–483.
- [9] CHEN, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3** 534–545.
- [10] CHEN, L. H. Y. AND SHAO, Q.-M. (2005)., Steins method for normal approximation. In *An introduction to Stein's method*, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. **4** 1–59.
- [11] CHEN, L. H. Y., GOLDSTEIN, L. AND SHAO, Q.-M. (2011). *Normal approximation by Stein's method*. Probability and its Applications (New York). Springer, Heidelberg.
- [12] DIACONIS, P. AND ZABELL, S. (1991). Closed form summation for classical distributions: variations on a theme of de Moivre. *Statist. Sci.* **6** 284–302.
- [13] EICHELSBACHER, P. AND LÖWE, M. (2010). Stein's method for dependent random variables occurring in statistical mechanics. *Electron. J. Probab.* **15** 962–988.

- [14] EICHELSBACHER, P. AND REINERT, G. (2008). Stein's method for discrete Gibbs measures. *Ann. Appl. Probab.* **18** 1588–1618.
- [15] ERHARDSSON, T. (2005). Stein's method for Poisson and compound Poisson approximation. In *An introduction to Stein's method*, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap. **4** 61–113.
- [16] GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** 724–739.
- [17] GÖTZE, F. AND TIKHOMIROV, A. N. (2006). Limit theorems for spectra of random matrices with martingale structure. *Teor. Veroyatn. Primen.* **51** 171–192; translation in *Theory Probab. Appl.* **51** 42–64 (2007).
- [18] GOLDSTEIN, L. AND RINOTT, Y. (1996). On multivariate normal approximations by Stein's method and size bias couplings. *J. Appl. Probab.* **33** 1–17.
- [19] HUDSON, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.* **6** 473–484.
- [20] HWANG, J. T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases. *Ann. Statist.* **10** 857–867.
- [21] JOHNSON, O. (2004). *Information theory and the central limit theorem*. Imperial College Press, London, UK.
- [22] JOHNSON, O. AND BARRON, A. B. (2004). Fisher information inequalities and the central limit theorem. *Probab. Theory Relat. Fields* **129** 391–409.
- [23] KONTOYIANNIS, I., HARREMOËS, P. AND JOHNSON, O. (2005). Entropy and the law of small numbers. *IEEE Transactions on information theory* **51** 466–472.
- [24] LEHMANN, E. L. AND CASELLA, G. (1998). *Theory of point estimation*. Second edition. Springer Texts in Statistics. Springer-Verlag, New York.
- [25] LEY, C. AND SWAN, Y. (2011). On a connection between Stein characterizations and Fisher information.
- [26] NOURDIN, I. AND PECCATI, G. (2010a). Stein's method meets Malliavin calculus: a short survey with new estimates. In *Recent Advances in Stochastic Dynamics and Stochastic Analysis*, World Scientific.
- [27] NOURDIN, I. AND PECCATI, G. (2011). Stein's method on Wiener chaos. *Probab. Theory Related Fields*. To appear.
- [28] NOURDIN, I., PECCATI, G. AND RÉVEILLAC, A. (2010b). Multivariate normal approximation using Stein's method and Malliavin calculus. *Ann. Inst. Henri Poincaré Probab. Stat.* **46** 45–58.
- [29] PEKÖZ, E., RÖLLIN, A. AND ROSS, N. Degree asymptotics with rates for preferential attachment random graphs.
- [30] REINERT, G. AND RÖLLIN, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Ann. Probab.* **37** 2150–2173.
- [31] ROSS, N. (2011). Fundamentals of Steins method *Probab. Surv.* **8** 210–293.
- [32] SCHOUTENS, W. (2001) Orthogonal Polynomials in Stein's Method. *J. Math. Anal. Appl.* **253** 515–531.
- [33] STEIN, C. (1972). A bound for the error in the normal approximation to

- the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. Math. Statist. Probab. Vol. II: Probability Theory* 583–602.
- [34] STEIN, C. (1986). *Approximate computation of expectations*. IMS Lecture Notes Monogr. Ser. **7**, IMS, Hayward, CA.
- [35] STEIN, C. WITH DIACONIS, P., HOLMES, S. AND REINERT, G. (2004). Use of exchangeable pairs in the analysis of simulations. In *Stein's Method: Expository Lectures and Applications*, *IMS Lecture Notes Monograph Series* **46** 69–77.